wrangle_act

September 14, 2020

0.0.1 Getting Started

Importing modules and downloading files

[1]:
```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import requests
import io

%matplotlib inline
sns.set_style('darkgrid')
```

[2]:
```
pd.set_option('display.max_colwidth', None)
```

[3]:
```
pulled = requests.get('https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
↪599fd2ad_image-predictions/image-predictions.tsv')

with open('image_predictions.txt', 'w') as f:
    f.write( pulled.text )
```

[4]:
```
tweets = pd.read_json('tweet-json copy', lines=True)
main = pd.read_csv('twitter-archive-enhanced.csv')
```

[5]:
```
predictions = pd.read_csv('image_predictions.txt', sep='\t')
predictions.sample(5)
```

[5]:
```
                tweet_id                                           jpg_url  \
1184  738883359779196928  https://pbs.twimg.com/media/CkEKe3QWYAAwoDy.jpg
1239  746872823977771008  https://pbs.twimg.com/media/Cl1s1p7WMAA44Vk.jpg
895   699468677801091073  https://pbs.twimg.com/media/CbTvNpoW0AEemnx.jpg
1467  778748913645780993  https://pbs.twimg.com/media/Cs6r_-kVIAALh1p.jpg
1859  842115215311396866  https://pbs.twimg.com/media/C6_LTCZWoAAKm_0.jpg

      img_num                         p1   p1_conf  p1_dog                p2  \
1184        2         Labrador_retriever  0.691137    True  golden_retriever
```

```
1239        1                  Pembroke  0.540201    True            beagle
895         3                  Pembroke  0.969400    True         Cardigan
1467        1  Staffordshire_bullterrier  0.351434    True            boxer
1859        1                      chow  0.293493    True      Newfoundland

       p2_conf  p2_dog                            p3   p3_conf  p3_dog
1184  0.195558    True       Chesapeake_Bay_retriever  0.019585    True
1239  0.207835    True              Italian_greyhound  0.043565    True
895   0.026059    True                       Chihuahua  0.003505    True
1467  0.201478    True  American_Staffordshire_terrier  0.142838    True
1859  0.181336    True                      schipperke  0.125152    True
```

[6]:
```
main.sample(5)
```

[6]:
```
                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
1902  674644256330530816                    NaN                  NaN
911   757597904299253760                    NaN                  NaN
1329  705898680587526145                    NaN                  NaN
662   790987426131050500                    NaN                  NaN
1168  721001180231503872                    NaN                  NaN

                      timestamp  \
1902  2015-12-09 17:38:19 +0000
911   2016-07-25 15:26:30 +0000
1329  2016-03-04 23:32:15 +0000
662   2016-10-25 18:44:32 +0000
1168  2016-04-15 15:44:11 +0000

  source  \
1902  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
911   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
1329  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
662   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
1168  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

                                                                text  \
1902                                                             When you
see sophomores in high school driving. 11/10 https://t.co/m6aC8d1Kzp
911                           RT @jon_hill987: @dog_rates There is a cunningly
disguised pupper here mate! 11/10 at least. https://t.co/7boff8zojZ
1329  Meet Max. He's a Fallopian Cephalopuff. Eyes are magical af. Lil dandruff
problem. No big deal 10/10 would still pet https://t.co/c67nUjwmFs
```

```
662                                                                  This is
Misty. She has a cowboy hat on her nose. 12/10 https://t.co/Eno0mypHIr
1168          This is Oliver. Bath time is upon him. His fear of the wetness
postpones his ultimate pupper destiny. 11/10 https://t.co/AFzzKqR4tT

      retweeted_status_id  retweeted_status_user_id  \
1902                  NaN                       NaN
911          7.575971e+17               280479778.0
1329                  NaN                       NaN
662                   NaN                       NaN
1168                  NaN                       NaN

     retweeted_status_timestamp  \
1902                        NaN
911   2016-07-25 15:23:28 +0000
1329                        NaN
662                         NaN
1168                        NaN

                                          expanded_urls  \
1902
https://twitter.com/dog_rates/status/674644256330530816/photo/1
911   https://twitter.com/jon_hill987/status/757597141099548672/photo/1,https://
twitter.com/jon_hill987/status/757597141099548672/photo/1
1329      https://twitter.com/dog_rates/status/705898680587526145/photo/1,https:
//twitter.com/dog_rates/status/705898680587526145/photo/1
662
https://twitter.com/dog_rates/status/790987426131050500/photo/1
1168
https://twitter.com/dog_rates/status/721001180231503872/photo/1

      rating_numerator  rating_denominator    name doggo floofer  pupper puppo
1902                11                  10    None  None    None    None  None
911                 11                  10    None  None    None  pupper  None
1329                10                  10     Max  None    None    None  None
662                 12                  10   Misty  None    None    None  None
1168                11                  10  Oliver  None    None  pupper  None
```

[7]:
```
tweets.sample(5)
```

[7]:
```
                     created_at                  id              id_str  \
929  2016-07-16 01:08:03+00:00  754120377874386944  754120377874386944
296  2017-03-02 01:20:01+00:00  837110210464448512  837110210464448512
1167 2016-04-15 01:26:47+00:00  720785406564900865  720785406564900864
642  2016-10-31 21:00:23+00:00  793195938047070209  793195938047070208
1949 2015-12-07 02:13:55+00:00  673686845050527744  673686845050527744
```

```
                                                  full_text  \
929   When you hear your owner say they need to hatch another egg, but you've
already been on 17 walks today. 10/10 https://t.co/lFEoGqZ4oA
296                                This is Clark. He passed pupper training
today. Round of appaws for Clark. 13/10 https://t.co/7pUjwe8X6B
1167                               This is Archie. He hears everything you say.
Doesn't matter where you are. 12/10 https://t.co/0l4I8famYp
642             Say hello to Lily. She's pupset that her costume doesn't fit as
well as last year. 12/10 poor puppo https://t.co/YSi6K1firY
1949                                           This is George. He's upset that
the 4th of July isn't everyday. 11/10 https://t.co/wImU0jdx3E

      truncated display_text_range  \
929       False           [0, 109]
296       False            [0, 80]
1167      False           [0, 104]
642       False            [0, 99]
1949      False            [0, 93]

                                    entities  \
929   {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [], 'media':
[{'id': 754120358878412800, 'id_str': '754120358878412800', 'indices': [110,
133], 'media_url': 'http://pbs.twimg.com/media/CncseIzWgAA4ghH.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CncseIzWgAA4ghH.jpg', 'url':
'https://t.co/lFEoGqZ4oA', 'display_url': 'pic.twitter.com/lFEoGqZ4oA',
'expanded_url':
'https://twitter.com/dog_rates/status/754120377874386944/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 383, 'h': 680, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 576, 'h': 1024, 'resize':
'fit'}, 'medium': {'w': 576, 'h': 1024, 'resize': 'fit'}}}]}
296   {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [], 'media':
[{'id': 837110202889617409, 'id_str': '837110202889617409', 'indices': [81,
104], 'media_url': 'http://pbs.twimg.com/media/C54DS1kXQAEU5pS.jpg',
'media_url_https': 'https://pbs.twimg.com/media/C54DS1kXQAEU5pS.jpg', 'url':
'https://t.co/7pUjwe8X6B', 'display_url': 'pic.twitter.com/7pUjwe8X6B',
'expanded_url':
'https://twitter.com/dog_rates/status/837110210464448512/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 545, 'h': 680, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'medium': {'w': 961, 'h': 1200,
'resize': 'fit'}, 'large': {'w': 1640, 'h': 2048, 'resize': 'fit'}}}]}
1167    {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [],
'media': [{'id': 720785400575455232, 'id_str': '720785400575455232', 'indices':
[81, 104], 'media_url': 'http://pbs.twimg.com/media/CgC-gMCWcAAawUE.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CgC-gMCWcAAawUE.jpg', 'url':
'https://t.co/0l4I8famYp', 'display_url': 'pic.twitter.com/0l4I8famYp',
'expanded_url':
'https://twitter.com/dog_rates/status/720785406564900865/photo/1', 'type':
```

'photo', 'sizes': {'small': {'w': 340, 'h': 453, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'medium': {'w': 600, 'h': 800, 'resize':
'fit'}, 'large': {'w': 768, 'h': 1024, 'resize': 'fit'}}}]}
642   {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [],
'media': [{'id': 793195928287055872, 'id_str': '793195928287055872', 'indices':
[100, 123], 'media_url': 'http://pbs.twimg.com/media/CwH_fobWEAAfJf8.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CwH_fobWEAAfJf8.jpg', 'url':
'https://t.co/YSi6K1firY', 'display_url': 'pic.twitter.com/YSi6K1firY',
'expanded_url':
'https://twitter.com/dog_rates/status/793195938047070209/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 639, 'h': 423, 'resize': 'fit'}, 'large':
{'w': 639, 'h': 423, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'small': {'w': 639, 'h': 423, 'resize': 'fit'}}}]}
1949   {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [],
'media': [{'id': 673686768529645568, 'id_str': '673686768529645568', 'indices':
[70, 93], 'media_url': 'http://pbs.twimg.com/media/CVlqi_AXIAASlcD.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CVlqi_AXIAASlcD.jpg', 'url':
'https://t.co/wImU0jdx3E', 'display_url': 'pic.twitter.com/wImU0jdx3E',
'expanded_url':
'https://twitter.com/dog_rates/status/673686845050527744/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 340, 'h': 604, 'resize': 'fit'}, 'large':
{'w': 576, 'h': 1024, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'medium': {'w': 576, 'h': 1024, 'resize': 'fit'}}}]}

   extended_entities \
929
{'media': [{'id': 754120358878412800, 'id_str': '754120358878412800', 'indices':
[110, 133], 'media_url': 'http://pbs.twimg.com/media/CncseIzWgAA4ghH.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CncseIzWgAA4ghH.jpg', 'url':
'https://t.co/lFEoGqZ4oA', 'display_url': 'pic.twitter.com/lFEoGqZ4oA',
'expanded_url':
'https://twitter.com/dog_rates/status/754120377874386944/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 383, 'h': 680, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 576, 'h': 1024, 'resize':
'fit'}, 'medium': {'w': 576, 'h': 1024, 'resize': 'fit'}}}]}
296
{'media': [{'id': 837110202889617409, 'id_str': '837110202889617409', 'indices':
[81, 104], 'media_url': 'http://pbs.twimg.com/media/C54DS1kXQAEU5pS.jpg',
'media_url_https': 'https://pbs.twimg.com/media/C54DS1kXQAEU5pS.jpg', 'url':
'https://t.co/7pUjwe8X6B', 'display_url': 'pic.twitter.com/7pUjwe8X6B',
'expanded_url':
'https://twitter.com/dog_rates/status/837110210464448512/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 545, 'h': 680, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'medium': {'w': 961, 'h': 1200,
'resize': 'fit'}, 'large': {'w': 1640, 'h': 2048, 'resize': 'fit'}}}]}
1167
{'media': [{'id': 720785400575455232, 'id_str': '720785400575455232', 'indices':

```
[81, 104], 'media_url': 'http://pbs.twimg.com/media/CgC-gMCWcAAawUE.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CgC-gMCWcAAawUE.jpg', 'url':
'https://t.co/0l4I8famYp', 'display_url': 'pic.twitter.com/0l4I8famYp',
'expanded_url':
'https://twitter.com/dog_rates/status/720785406564900865/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 340, 'h': 453, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'medium': {'w': 600, 'h': 800, 'resize':
'fit'}, 'large': {'w': 768, 'h': 1024, 'resize': 'fit'}}}]}
642   {'media': [{'id': 793195928287055872, 'id_str': '793195928287055872',
'indices': [100, 123], 'media_url':
'http://pbs.twimg.com/media/CwH_fobWEAAfJf8.jpg', 'media_url_https':
'https://pbs.twimg.com/media/CwH_fobWEAAfJf8.jpg', 'url':
'https://t.co/YSi6K1firY', 'display_url': 'pic.twitter.com/YSi6K1firY',
'expanded_url':
'https://twitter.com/dog_rates/status/793195938047070209/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 639, 'h': 423, 'resize': 'fit'}, 'large':
{'w': 639, 'h': 423, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'small': {'w': 639, 'h': 423, 'resize': 'fit'}}}, {'id':
793195928274501633, 'id_str': '793195928274501633', 'indices': [100, 123],
'media_url': 'http://pbs.twimg.com/media/CwH_foYWgAEvTyI.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CwH_foYWgAEvTyI.jpg', 'url':
'https://t.co/YSi6K1firY', 'display_url': 'pic.twitter.com/YSi6K1firY',
'expanded_url':
'https://twitter.com/dog_rates/status/793195938047070209/photo/1', 'type':
'photo', 'sizes': {'large': {'w': 2048, 'h': 1530, 'resize': 'fit'}, 'medium':
{'w': 1200, 'h': 896, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'small': {'w': 680, 'h': 508, 'resize': 'fit'}}}]}
1949
{'media': [{'id': 673686768529645568, 'id_str': '673686768529645568', 'indices':
[70, 93], 'media_url': 'http://pbs.twimg.com/media/CVlqi_AXIAASlcD.jpg',
'media_url_https': 'https://pbs.twimg.com/media/CVlqi_AXIAASlcD.jpg', 'url':
'https://t.co/wImU0jdx3E', 'display_url': 'pic.twitter.com/wImU0jdx3E',
'expanded_url':
'https://twitter.com/dog_rates/status/673686845050527744/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 340, 'h': 604, 'resize': 'fit'}, 'large':
{'w': 576, 'h': 1024, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'medium': {'w': 576, 'h': 1024, 'resize': 'fit'}}}]}

  source  \
929   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
296   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
1167  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
642   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
```

```
1949  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

      in_reply_to_status_id  …  favorite_count  favorited  retweeted  \
929                     NaN  …            8655      False      False
296                     NaN  …           17480      False      False
1167                    NaN  …            3400      False      False
642                     NaN  …           17063      False      False
1949                    NaN  …            1544      False      False

     possibly_sensitive possibly_sensitive_appealable  lang  retweeted_status  \
929                 0.0                          0.0    en               NaN
296                 0.0                          0.0    en               NaN
1167                0.0                          0.0    en               NaN
642                 0.0                          0.0    en               NaN
1949                0.0                          0.0    en               NaN

     quoted_status_id  quoted_status_id_str  quoted_status
929               NaN                   NaN            NaN
296               NaN                   NaN            NaN
1167              NaN                   NaN            NaN
642               NaN                   NaN            NaN
1949              NaN                   NaN            NaN

[5 rows x 31 columns]
```

Creating a function to drop columns with less than 2 values

[8]:
```
def drop_useless(dataframe):
    useless = []
    for i in dataframe.columns:
        try:
            if(dataframe[i].nunique()<2):
                useless.append(i)
        except TypeError:
            pass
    print('\nremoved columns: {}\n\n'.format(useless))
    return dataframe.drop(useless, axis=1)
```

[9]:
```
df = drop_useless(tweets)
df.head()
```

```
removed columns: ['truncated', 'geo', 'coordinates', 'contributors',
'retweeted', 'possibly_sensitive', 'possibly_sensitive_appealable']
```

[9]:
```
                  created_at                   id              id_str  \
0 2017-08-01 16:23:56+00:00  892420643555336193  892420643555336192
1 2017-08-01 00:17:27+00:00  892177421306343426  892177421306343424
2 2017-07-31 00:18:03+00:00  891815181378084864  891815181378084864
3 2017-07-30 15:58:51+00:00  891689557279858688  891689557279858688
4 2017-07-29 16:00:24+00:00  891327558926688256  891327558926688256

                                                full_text  \
0                                                          This is Phineas. He's a
mystical boy. Only ever appears in the hole of a donut. 13/10
https://t.co/MgUWQ76dJU
1  This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not,
she's available for pats, snugs, boops, the whole bit. 13/10
https://t.co/0Xxu71qeIV
2                   This is Archie. He is a rare Norwegian Pouncing Corgo. Lives
in the tall grass. You never know when one may strike. 12/10
https://t.co/wUnZnhtVJB
3                                                                  This is Darla. She
commenced a snooze mid meal. 13/10 happens to the best of us
https://t.co/tD36da7qLQ
4  This is Franklin. He would like you to stop calling him "cute." He is a very
fierce shark and should be respected as such. 12/10 #BarkWeek
https://t.co/AtUZn91f7f

  display_text_range  \
0            [0, 85]
1           [0, 138]
2           [0, 121]
3            [0, 79]
4           [0, 138]

                                                entities  \
0                                                {'hashtags': [], 'symbols': [],
'user_mentions': [], 'urls': [], 'media': [{'id': 892420639486877696, 'id_str':
'892420639486877696', 'indices': [86, 109], 'media_url':
'http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg', 'media_url_https':
'https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg', 'url':
'https://t.co/MgUWQ76dJU', 'display_url': 'pic.twitter.com/MgUWQ76dJU',
'expanded_url':
'https://twitter.com/dog_rates/status/892420643555336193/photo/1', 'type':
'photo', 'sizes': {'large': {'w': 540, 'h': 528, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 540, 'h': 528, 'resize':
'fit'}, 'medium': {'w': 540, 'h': 528, 'resize': 'fit'}}}]}
1                                          {'hashtags': [], 'symbols': [],
'user_mentions': [], 'urls': [], 'media': [{'id': 892177413194625024, 'id_str':
'892177413194625024', 'indices': [139, 162], 'media_url':
'http://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg', 'media_url_https':
```

'https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg', 'url':
'https://t.co/0Xxu71qeIV', 'display_url': 'pic.twitter.com/0Xxu71qeIV',
'expanded_url':
'https://twitter.com/dog_rates/status/892177421306343426/photo/1', 'type':
'photo', 'sizes': {'large': {'w': 1407, 'h': 1600, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 598, 'h': 680, 'resize':
'fit'}, 'medium': {'w': 1055, 'h': 1200, 'resize': 'fit'}}}]}
2 {'hashtags': [], 'symbols': [],
'user_mentions': [], 'urls': [], 'media': [{'id': 891815175371796480, 'id_str':
'891815175371796480', 'indices': [122, 145], 'media_url':
'http://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg', 'media_url_https':
'https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg', 'url':
'https://t.co/wUnZnhtVJB', 'display_url': 'pic.twitter.com/wUnZnhtVJB',
'expanded_url':
'https://twitter.com/dog_rates/status/891815181378084864/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 901, 'h': 1200, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 1201, 'h': 1600,
'resize': 'fit'}, 'small': {'w': 510, 'h': 680, 'resize': 'fit'}}}]}
3 {'hashtags': [], 'symbols': [],
'user_mentions': [], 'urls': [], 'media': [{'id': 891689552724799489, 'id_str':
'891689552724799489', 'indices': [80, 103], 'media_url':
'http://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg', 'media_url_https':
'https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg', 'url':
'https://t.co/tD36da7qLQ', 'display_url': 'pic.twitter.com/tD36da7qLQ',
'expanded_url':
'https://twitter.com/dog_rates/status/891689557279858688/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 901, 'h': 1200, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 1201, 'h': 1600,
'resize': 'fit'}, 'small': {'w': 510, 'h': 680, 'resize': 'fit'}}}]}
4 {'hashtags': [{'text': 'BarkWeek', 'indices': [129, 138]}], 'symbols': [],
'user_mentions': [], 'urls': [], 'media': [{'id': 891327551943041024, 'id_str':
'891327551943041024', 'indices': [139, 162], 'media_url':
'http://pbs.twimg.com/media/DF6hr6AVYAAZ8G8.jpg', 'media_url_https':
'https://pbs.twimg.com/media/DF6hr6AVYAAZ8G8.jpg', 'url':
'https://t.co/AtUZn91f7f', 'display_url': 'pic.twitter.com/AtUZn91f7f',
'expanded_url':
'https://twitter.com/dog_rates/status/891327558926688256/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 680, 'h': 510, 'resize': 'fit'}, 'large':
{'w': 720, 'h': 540, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'medium': {'w': 720, 'h': 540, 'resize': 'fit'}}}]}

extended_entities \
0
{'media': [{'id': 892420639486877696, 'id_str': '892420639486877696', 'indices':
[86, 109], 'media_url': 'http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg',
'media_url_https': 'https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg', 'url':
'https://t.co/MgUWQ76dJU', 'display_url': 'pic.twitter.com/MgUWQ76dJU',

'expanded_url':
'https://twitter.com/dog_rates/status/892420643555336193/photo/1', 'type':
'photo', 'sizes': {'large': {'w': 540, 'h': 528, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 540, 'h': 528, 'resize':
'fit'}, 'medium': {'w': 540, 'h': 528, 'resize': 'fit'}}}]}
1
{'media': [{'id': 892177413194625024, 'id_str': '892177413194625024', 'indices':
[139, 162], 'media_url': 'http://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg',
'media_url_https': 'https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg', 'url':
'https://t.co/0Xxu71qeIV', 'display_url': 'pic.twitter.com/0Xxu71qeIV',
'expanded_url':
'https://twitter.com/dog_rates/status/892177421306343426/photo/1', 'type':
'photo', 'sizes': {'large': {'w': 1407, 'h': 1600, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 598, 'h': 680, 'resize':
'fit'}, 'medium': {'w': 1055, 'h': 1200, 'resize': 'fit'}}}]}
2
{'media': [{'id': 891815175371796480, 'id_str': '891815175371796480', 'indices':
[122, 145], 'media_url': 'http://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg',
'media_url_https': 'https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg', 'url':
'https://t.co/wUnZnhtVJB', 'display_url': 'pic.twitter.com/wUnZnhtVJB',
'expanded_url':
'https://twitter.com/dog_rates/status/891815181378084864/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 901, 'h': 1200, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 1201, 'h': 1600,
'resize': 'fit'}, 'small': {'w': 510, 'h': 680, 'resize': 'fit'}}}]}
3
{'media': [{'id': 891689552724799489, 'id_str': '891689552724799489', 'indices':
[80, 103], 'media_url': 'http://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg',
'media_url_https': 'https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg', 'url':
'https://t.co/tD36da7qLQ', 'display_url': 'pic.twitter.com/tD36da7qLQ',
'expanded_url':
'https://twitter.com/dog_rates/status/891689557279858688/photo/1', 'type':
'photo', 'sizes': {'medium': {'w': 901, 'h': 1200, 'resize': 'fit'}, 'thumb':
{'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 1201, 'h': 1600,
'resize': 'fit'}, 'small': {'w': 510, 'h': 680, 'resize': 'fit'}}}]}
4 {'media': [{'id': 891327551943041024, 'id_str': '891327551943041024',
'indices': [139, 162], 'media_url':
'http://pbs.twimg.com/media/DF6hr6AVYAAZ8G8.jpg', 'media_url_https':
'https://pbs.twimg.com/media/DF6hr6AVYAAZ8G8.jpg', 'url':
'https://t.co/AtUZn91f7f', 'display_url': 'pic.twitter.com/AtUZn91f7f',
'expanded_url':
'https://twitter.com/dog_rates/status/891327558926688256/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 680, 'h': 510, 'resize': 'fit'}, 'large':
{'w': 720, 'h': 540, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'medium': {'w': 720, 'h': 540, 'resize': 'fit'}}}, {'id':
891327551947157504, 'id_str': '891327551947157504', 'indices': [139, 162],
'media_url': 'http://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg',

```
'media_url_https': 'https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg', 'url':
'https://t.co/AtUZn91f7f', 'display_url': 'pic.twitter.com/AtUZn91f7f',
'expanded_url':
'https://twitter.com/dog_rates/status/891327558926688256/photo/1', 'type':
'photo', 'sizes': {'small': {'w': 680, 'h': 510, 'resize': 'fit'}, 'large':
{'w': 720, 'h': 540, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize':
'crop'}, 'medium': {'w': 720, 'h': 540, 'resize': 'fit'}}}]}

source  \
0  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
1  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
2  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
3  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
4  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

   in_reply_to_status_id  in_reply_to_status_id_str  …  place  \
0                    NaN                        NaN  …   None
1                    NaN                        NaN  …   None
2                    NaN                        NaN  …   None
3                    NaN                        NaN  …   None
4                    NaN                        NaN  …   None

   is_quote_status retweet_count favorite_count favorited  lang  \
0            False          8853          39467     False    en
1            False          6514          33819     False    en
2            False          4328          25461     False    en
3            False          8964          42908     False    en
4            False          9774          41048     False    en

   retweeted_status  quoted_status_id  quoted_status_id_str quoted_status
0               NaN               NaN                   NaN           NaN
1               NaN               NaN                   NaN           NaN
2               NaN               NaN                   NaN           NaN
3               NaN               NaN                   NaN           NaN
4               NaN               NaN                   NaN           NaN

[5 rows x 24 columns]
```

[10]:
```
archive = drop_useless(main)
archive.sample(5)
```

```
removed columns: []
```

[10]:

```
                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
1805  676942428000112642                    NaN                  NaN
2253  667793409583771648                    NaN                  NaN
739   780601303617732608                    NaN                  NaN
2343  666073100786774016                    NaN                  NaN
87    875144289856114688                    NaN                  NaN

                      timestamp  \
1805  2015-12-16 01:50:26 +0000
2253  2015-11-20 19:55:30 +0000
739   2016-09-27 02:53:48 +0000
2343  2015-11-16 01:59:36 +0000
87    2017-06-15 00:13:52 +0000

  source  \
1805  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
2253  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
739   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
2343  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
87    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

text  \
1805
Who leaves the last cupcake just sitting there? 9/10 https://t.co/PWMqAoEx2a
2253                                          Dogs only please. Small cows and
other non canines will not be tolerated. Sick tattoos tho 8/10
https://t.co/s1z7mX4c9O
739                                    Meet Hercules. He can have whatever he
wants for the rest of eternity. 12/10 would snug passionately
https://t.co/mH0IOyFdIG
2343                       Let's hope this flight isn't Malaysian (lol). What a
dog! Almost completely camouflaged. 10/10 I trust this pilot
https://t.co/Yk6GHE9tOY
87    Meet Nugget and Hank. Nugget took Hank's bone. Hank is wondering if you
would please return it to him. Both 13/10 would not intervene
https://t.co/ogith9ejNj

      retweeted_status_id  retweeted_status_user_id  \
```

```
1805                  NaN                         NaN
2253                  NaN                         NaN
739                   NaN                         NaN
2343                  NaN                         NaN
87                    NaN                         NaN

     retweeted_status_timestamp  \
1805                        NaN
2253                        NaN
739                         NaN
2343                        NaN
87                          NaN

                                                    expanded_urls  \
1805  https://twitter.com/dog_rates/status/676942428000112642/photo/1
2253  https://twitter.com/dog_rates/status/667793409583771648/photo/1
739   https://twitter.com/dog_rates/status/780601303617732608/photo/1
2343  https://twitter.com/dog_rates/status/666073100786774016/photo/1
87    https://twitter.com/dog_rates/status/875144289856114688/video/1

      rating_numerator  rating_denominator      name doggo floofer pupper  \
1805                 9                  10      None  None    None   None
2253                 8                  10      None  None    None   None
739                 12                  10  Hercules  None    None   None
2343                10                  10      None  None    None   None
87                  13                  10    Nugget  None    None   None

     puppo
1805  None
2253  None
739   None
2343  None
87    None
```

[11]:
```
predict = drop_useless(predictions)
predict.query('p1_dog == False')
```

```
removed columns: []
```

[11]:
```
                tweet_id  \
6     666051853826850816
8     666057090499244032
17    666104133288665088
18    666268910803644416
```

```
21    666293911632134144
…                  …
2026  882045870035918850
2046  886680336477933568
2052  887517139158093824
2071  891689557279858688
2074  892420643555336193

      jpg_url  \
6
https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg
8
https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg
17
https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg
18
https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg
21
https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg
…
…
2026
https://pbs.twimg.com/media/DD2oCl2WAAEI_4a.jpg
2046
https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg
2052  https://pbs.twimg.com/ext_tw_video_thumb/887517108413886465/pu/img/WanJKws
sZj4VJvL9.jpg
2071
https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg
2074
https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg

      img_num                p1   p1_conf  p1_dog                   p2  \
6           1        box_turtle  0.933012   False          mud_turtle
8           1     shopping_cart  0.962465   False     shopping_basket
17          1               hen  0.965932   False                cock
18          1  desktop_computer  0.086502   False                desk
21          1  three-toed_sloth  0.914671   False               otter
…       …                …       …     …                   …
2026        1          web_site  0.949591   False               dhole
2046        1       convertible  0.738995   False          sports_car
2052        1         limousine  0.130432   False           tow_truck
2071        1       paper_towel  0.170278   False  Labrador_retriever
2074        1            orange  0.097049   False               bagel

       p2_conf  p2_dog               p3   p3_conf  p3_dog
6     0.045885   False          terrapin  0.017885   False
```

```
8     0.014594   False  golden_retriever  0.007959    True
17    0.033919   False         partridge  0.000052   False
18    0.085547   False          bookcase  0.079480   False
21    0.015250   False    great_grey_owl  0.013207   False
…          …     …                …       …        …
2026  0.017326   False  golden_retriever  0.006941    True
2046  0.139952   False         car_wheel  0.044173   False
2052  0.029175   False     shopping_cart  0.026321   False
2071  0.168086    True           spatula  0.040836   False
2074  0.085851   False            banana  0.076110   False

[543 rows x 12 columns]
```

[12]:
```
print(predict.dtypes, archive.dtypes, df.dtypes, sep='\n\n=========\n\n')
```

```
tweet_id      int64
jpg_url      object
img_num       int64
p1           object
p1_conf     float64
p1_dog         bool
p2           object
p2_conf     float64
p2_dog         bool
p3           object
p3_conf     float64
p3_dog         bool
dtype: object

=========

tweet_id                      int64
in_reply_to_status_id       float64
in_reply_to_user_id         float64
timestamp                    object
source                       object
text                         object
retweeted_status_id         float64
retweeted_status_user_id    float64
retweeted_status_timestamp   object
expanded_urls                object
rating_numerator              int64
rating_denominator            int64
name                         object
doggo                        object
floofer                      object
pupper                       object
```

```
puppo                          object
dtype: object

=========

created_at                  datetime64[ns, UTC]
id                                        int64
id_str                                    int64
full_text                                object
display_text_range                       object
entities                                 object
extended_entities                        object
source                                   object
in_reply_to_status_id                   float64
in_reply_to_status_id_str               float64
in_reply_to_user_id                     float64
in_reply_to_user_id_str                 float64
in_reply_to_screen_name                  object
user                                     object
place                                    object
is_quote_status                            bool
retweet_count                             int64
favorite_count                            int64
favorited                                  bool
lang                                     object
retweeted_status                         object
quoted_status_id                        float64
quoted_status_id_str                    float64
quoted_status                            object
dtype: object
```

converting dates to appropiate format

[13]:
```
archive['timestamp'] = pd.to_datetime(archive['timestamp'])
archive['timestamp']
```

[13]:
```
0      2017-08-01 16:23:56+00:00
1      2017-08-01 00:17:27+00:00
2      2017-07-31 00:18:03+00:00
3      2017-07-30 15:58:51+00:00
4      2017-07-29 16:00:24+00:00
                  …
2351   2015-11-16 00:24:50+00:00
2352   2015-11-16 00:04:52+00:00
2353   2015-11-15 23:21:54+00:00
2354   2015-11-15 23:05:30+00:00
2355   2015-11-15 22:32:08+00:00
Name: timestamp, Length: 2356, dtype: datetime64[ns, UTC]
```

extracting source name and url.

[14]:
```
source_url = archive['source'].apply(lambda x:x.split('"')[1])
archive['source'] = archive['source'].apply(lambda x:x.split('"')[-1][1:-4])
archive['source_url'] = source_url
```

[15]:
```
col = list(archive.columns)
col.insert(5, col.pop(-1))
archive = archive[col]
```

Converting 'None' strings to Null Values

[16]:
```
#convert 'None' Strings to Null values
archive.replace('None', np.nan, inplace=True)
```

[17]:
```
archive.shape
```

[17]:
```
(2356, 18)
```

[18]:
```
predict.shape
```

[18]:
```
(2075, 12)
```

[19]:
```
type(archive['name'].value_counts().index)
```

[19]:
```
pandas.core.indexes.base.Index
```

Excluding names that are lower case, which all happen to be like: (a, an, this, the, etc...)

[20]:
```
#set wrong names to nan
wrong_names = list(archive['name'].astype(str).str.islower())
archive.loc[wrong_names, 'name'] = np.nan
```

Removing data about retweets from other users, as only original tweets data was requested.

[21]:
```
#remove retweets
archive.drop(archive[~archive.retweeted_status_id.isna()].index, inplace=True)
archive.drop(archive[~archive.in_reply_to_user_id.isna()].index, inplace=True)
```

Combining dog types into 1 column, some rows contain more than one type, the former is used.

[22]:
```
dog_types = archive[['doggo','floofer', 'puppo', 'pupper']]
dog_types.count().sum()
```

[22]:
```
347
```

[23]:
```
dada = dog_types['doggo'].combine(dog_types['floofer'], lambda x,y: y if pd.
↪isna(x) else x )
dada = dada.combine(dog_types['pupper'], lambda x,y: y if pd.isna(x) else x )
```

```
dada = dada.combine(dog_types['puppo'], lambda x,y: y if pd.isna(x) else x )
```

[24]:
```
dada.value_counts().sum()
```

[24]:
```
336
```

Dropping unuseful data.

[25]:
```
archive['type'] = dada
archive.drop(['doggo', 'floofer', 'pupper', 'puppo', 'retweeted_status_id',␣
↪'retweeted_status_user_id', 'retweeted_status_timestamp',␣
↪'in_reply_to_user_id', 'in_reply_to_status_id'], axis=1, inplace=True)
```

[26]:
```
archive.dtypes
```

[26]:
```
tweet_id                             int64
timestamp              datetime64[ns, UTC]
source                              object
source_url                          object
text                                object
expanded_urls                       object
rating_numerator                     int64
rating_denominator                   int64
name                                object
type                                object
dtype: object
```

[27]:
```
archive['rating_denominator'].value_counts()
```

[27]:
```
10     2080
50        3
11        2
80        2
7         1
170       1
150       1
120       1
110       1
90        1
70        1
40        1
20        1
2         1
Name: rating_denominator, dtype: int64
```

Combining the predictions dataset with the tweets data.

[28]:
```
combined = archive.merge(predict, on='tweet_id', how ='inner')
combined.shape
```

[28]:
```
(1971, 21)
```

[29]:
```
predict.shape
```

[29]:
```
(2075, 12)
```

[30]:
```
archive.shape
```

[30]:
```
(2097, 10)
```

[31]:
```
combined.head()
```

[31]:

	tweet_id	timestamp	source	source_url	text	expanded_urls
0	892420643555336193	2017-08-01 16:23:56+00:00	Twitter for iPhone	http://twitter.com/download/iphone	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	
1	892177421306343426	2017-08-01 00:17:27+00:00	Twitter for iPhone	http://twitter.com/download/iphone	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10 https://t.co/0Xxu71qeIV	
2	891815181378084864	2017-07-31 00:18:03+00:00	Twitter for iPhone	http://twitter.com/download/iphone	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co/wUnZnhtVJB	
3	891689557279858688	2017-07-30 15:58:51+00:00	Twitter for iPhone	http://twitter.com/download/iphone	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	
4	891327558926688256	2017-07-29 16:00:24+00:00	Twitter for iPhone	http://twitter.com/download/iphone	This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such. 12/10 #BarkWeek https://t.co/AtUZn91f7f	

```
0
https://twitter.com/dog_rates/status/892420643555336193/photo/1
1
https://twitter.com/dog_rates/status/892177421306343426/photo/1
2
https://twitter.com/dog_rates/status/891815181378084864/photo/1
3
https://twitter.com/dog_rates/status/891689557279858688/photo/1
4  https://twitter.com/dog_rates/status/891327558926688256/photo/1,https://twitt
er.com/dog_rates/status/891327558926688256/photo/1

   rating_numerator  rating_denominator      name type  … img_num  \
0                13                  10   Phineas  NaN  …       1
1                13                  10     Tilly  NaN  …       1
2                12                  10    Archie  NaN  …       1
3                13                  10     Darla  NaN  …       1
4                12                  10  Franklin  NaN  …       2

            p1   p1_conf  p1_dog                  p2   p2_conf  p2_dog  \
0       orange  0.097049   False               bagel  0.085851   False
1    Chihuahua  0.323581    True            Pekinese  0.090647    True
2    Chihuahua  0.716012    True            malamute  0.078253    True
3  paper_towel  0.170278   False  Labrador_retriever  0.168086    True
4       basset  0.555712    True    English_springer  0.225770    True

                          p3   p3_conf  p3_dog
0                     banana  0.076110   False
1                   papillon  0.068957    True
2                     kelpie  0.031379    True
3                    spatula  0.040836   False
4  German_short-haired_pointer  0.175219    True

[5 rows x 21 columns]
```

Dropping data rows, in which the first prediction was not a dog. keeping in mind that $p1\_conf > p2\_conf > p3\_conf$ so even if there are other predictions that say it is a dog, the certainty of such prediction would be less that the certainty of the first prediction which says it is not a dog

[32]:
```
combined.drop(combined[~combined['p1_dog']].index, inplace=True)
```

[33]:
```
combined.shape
```

[33]:
```
(1463, 21)
```

Merging the retweet and favorite counts from the twitter api based on the time of the tweets. because i was uncertain which column related to the tweet id, whether 'id' or id_str'

[34]:
```
semi_final = combined.merge(df[['created_at', 'retweet_count',␣
↪'favorite_count']], left_on='timestamp', right_on='created_at', how='left')
```

[35]:
```
semi_final.dtypes
```

[35]:
```
tweet_id                            int64
timestamp              datetime64[ns, UTC]
source                              object
source_url                          object
text                                object
expanded_urls                       object
rating_numerator                     int64
rating_denominator                   int64
name                                object
type                                object
jpg_url                             object
img_num                              int64
p1                                  object
p1_conf                            float64
p1_dog                                bool
p2                                  object
p2_conf                            float64
p2_dog                                bool
p3                                  object
p3_conf                            float64
p3_dog                                bool
created_at             datetime64[ns, UTC]
retweet_count                        int64
favorite_count                       int64
dtype: object
```

Dropping more columns which would not be useful in the analysis process.

[36]:
```
final = semi_final.drop(['source_url', 'text', 'expanded_urls', 'created_at',␣
↪'jpg_url'], axis=1)
```

[37]:
```
final.head()
```

[37]:

	tweet_id	timestamp	source	\
0	892177421306343426	2017-08-01 00:17:27+00:00	Twitter for iPhone	
1	891815181378084864	2017-07-31 00:18:03+00:00	Twitter for iPhone	
2	891327558926688256	2017-07-29 16:00:24+00:00	Twitter for iPhone	
3	891087950875897856	2017-07-29 00:08:17+00:00	Twitter for iPhone	
4	890971913173991426	2017-07-28 16:27:12+00:00	Twitter for iPhone	

	rating_numerator	rating_denominator	name	type	img_num	\
0	13	10	Tilly	NaN	1	

```
1                12                  10     Archie  NaN        1
2                12                  10   Franklin  NaN        2
3                13                  10        NaN  NaN        1
4                13                  10        Jax  NaN        1

                        p1   p1_conf  p1_dog                p2   p2_conf  \
0                 Chihuahua  0.323581    True          Pekinese  0.090647
1                 Chihuahua  0.716012    True          malamute  0.078253
2                    basset  0.555712    True  English_springer  0.225770
3  Chesapeake_Bay_retriever  0.425595    True     Irish_terrier  0.116317
4              Appenzeller  0.341703    True     Border_collie  0.199287

   p2_dog                          p3   p3_conf  p3_dog  retweet_count  \
0    True                    papillon  0.068957    True           6514
1    True                      kelpie  0.031379    True           4328
2    True  German_short-haired_pointer  0.175219    True           9774
3    True              Indian_elephant  0.076902   False           3261
4    True                    ice_lolly  0.193548   False           2158

   favorite_count
0           33819
1           25461
2           41048
3           20562
4           12041
```

Combining the ratings of the dogs into 1 column

[38]:
```
rating = final['rating_numerator'] / final['rating_denominator']
rating.value_counts()
```

[38]:
```
1.200000    380
1.000000    320
1.100000    310
1.300000    208
0.900000    107
0.800000     58
0.700000     24
1.400000     19
0.600000     12
0.500000     11
0.400000      4
0.300000      3
0.200000      2
2.600000      1
2.700000      1
3.428571      1
```

```
0.818182      1
7.500000      1
dtype: int64
```

[39]:
```
final.drop(['rating_denominator'], axis = 1, inplace=True)
final.loc[:,'rating_numerator'] = rating
final.rename(columns={'rating_numerator': 'rating'}, inplace=True)
final.head()
```

[39]:

	tweet_id	timestamp	source	rating
0	892177421306343426	2017-08-01 00:17:27+00:00	Twitter for iPhone	1.3
1	891815181378084864	2017-07-31 00:18:03+00:00	Twitter for iPhone	1.2
2	891327558926688256	2017-07-29 16:00:24+00:00	Twitter for iPhone	1.2
3	891087950875897856	2017-07-29 00:08:17+00:00	Twitter for iPhone	1.3
4	890971913173991426	2017-07-28 16:27:12+00:00	Twitter for iPhone	1.3

	name	type	img_num	p1	p1_conf	p1_dog
0	Tilly	NaN	1	Chihuahua	0.323581	True
1	Archie	NaN	1	Chihuahua	0.716012	True
2	Franklin	NaN	2	basset	0.555712	True
3	NaN	NaN	1	Chesapeake_Bay_retriever	0.425595	True
4	Jax	NaN	1	Appenzeller	0.341703	True

	p2	p2_conf	p2_dog	p3	p3_conf
0	Pekinese	0.090647	True	papillon	0.068957
1	malamute	0.078253	True	kelpie	0.031379
2	English_springer	0.225770	True	German_short-haired_pointer	0.175219
3	Irish_terrier	0.116317	True	Indian_elephant	0.076902
4	Border_collie	0.199287	True	ice_lolly	0.193548

	p3_dog	retweet_count	favorite_count
0	True	6514	33819
1	True	4328	25461
2	True	9774	41048
3	False	3261	20562
4	False	2158	12041

Saving the data to A CSV file as requested.

[40]:
```
final.to_csv('twitter_archive_master.csv', index=False)
```

[41]:
```
df = pd.read_csv('twitter_archive_master.csv')
```

Reconverting the timestamp data into appropiate format as it was saved as a string

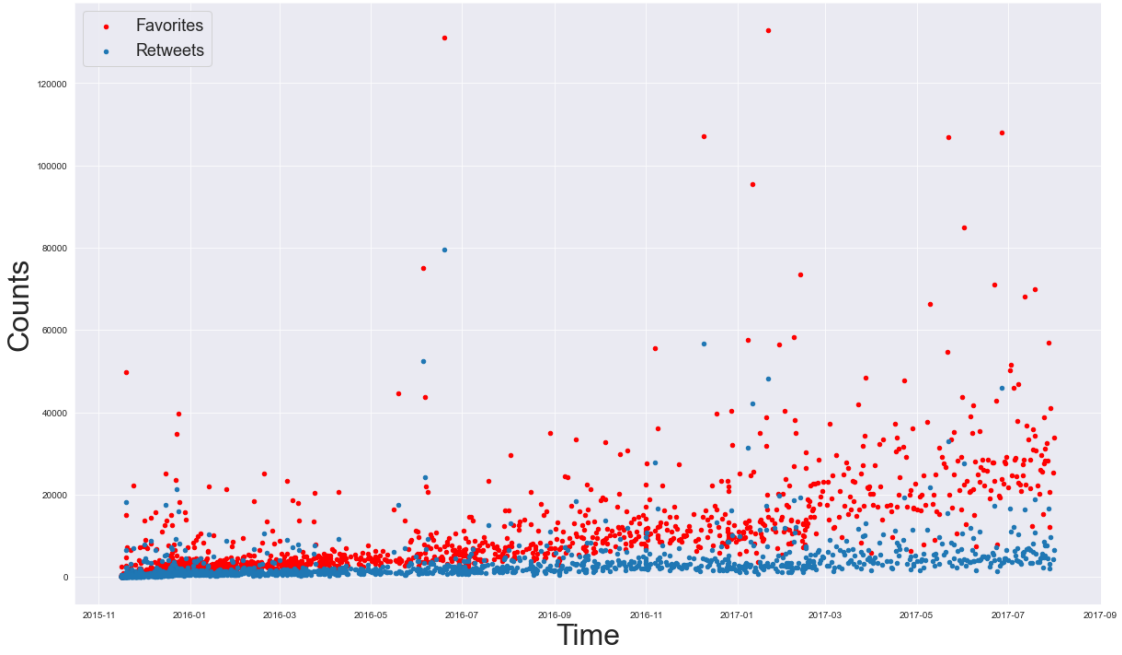
[42]:
```
df.loc[:,'timestamp'] = pd.to_datetime(df['timestamp'])
```

[43]:
```
df.dtypes
```

[43]:
```
tweet_id                            int64
timestamp          datetime64[ns, UTC]
source                             object
rating                            float64
name                               object
type                               object
img_num                             int64
p1                                 object
p1_conf                           float64
p1_dog                               bool
p2                                 object
p2_conf                           float64
p2_dog                               bool
p3                                 object
p3_conf                           float64
p3_dog                               bool
retweet_count                       int64
favorite_count                      int64
dtype: object
```

Studying the reaction with the tweets over time‘

[44]:
```
ax = df.plot(x='timestamp', y='favorite_count', kind='scatter',␣
 ↪figsize=(20,12), color='r');
df.plot(x='timestamp', y='retweet_count', kind='scatter', ax=ax);
ax.set_xlabel('Time', fontsize=32)
ax.set_ylabel('Counts', fontsize=32)
ax.legend(['Favorites', 'Retweets'], fontsize=18, loc=2);
```

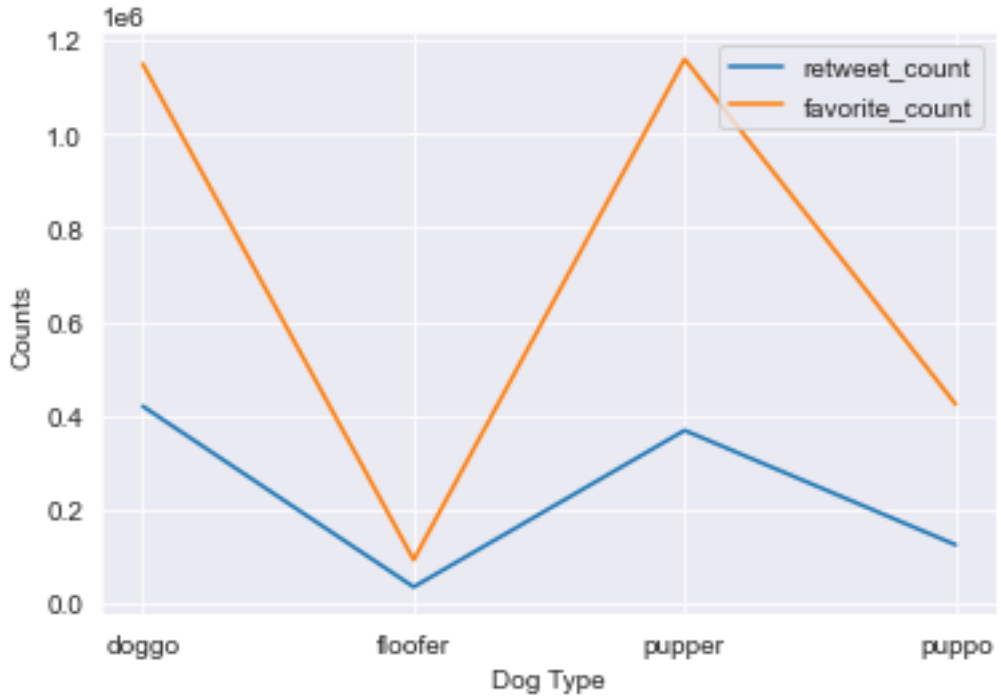


0.1 Results

Apparantly the favorite count almost doubled over the span of two years, in a sort of squared correlation. yet the retweet count is fairly the same, which tells us that the tweets need to be more engaging with the users.

Studying the correlation between dog type and interactions.

[45]:
```
types = df.groupby('type')[['retweet_count', 'favorite_count']].sum().
↪reset_index()
ax = types.plot()
ax.set_xticks([0,1,2,3])
ax.set_xticklabels(types['type'])
ax.set_xlabel('Dog Type')
ax.set_ylabel('Counts');
```

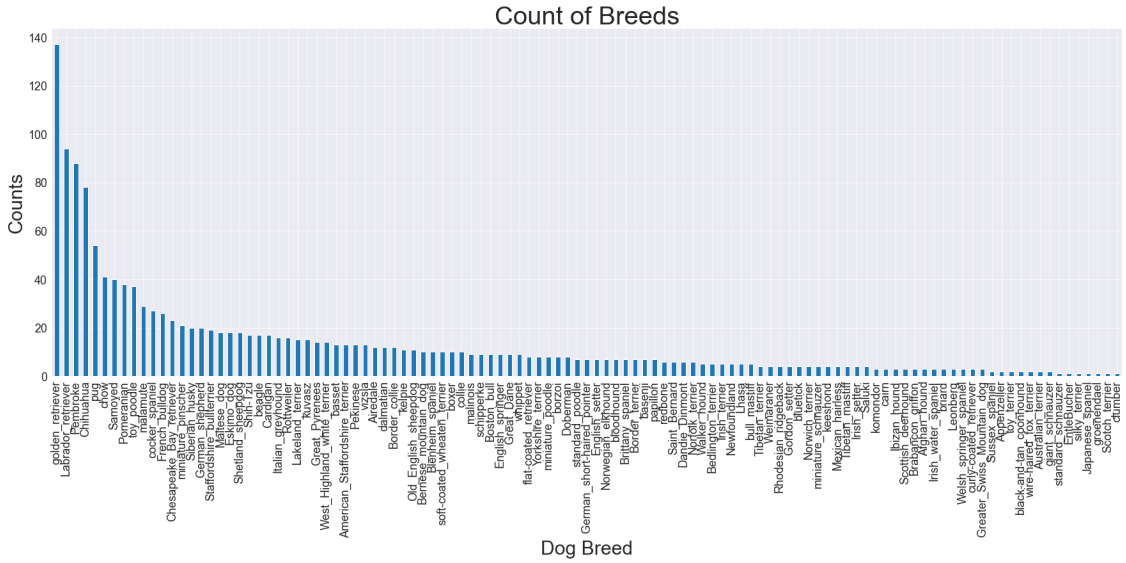


0.2 Results

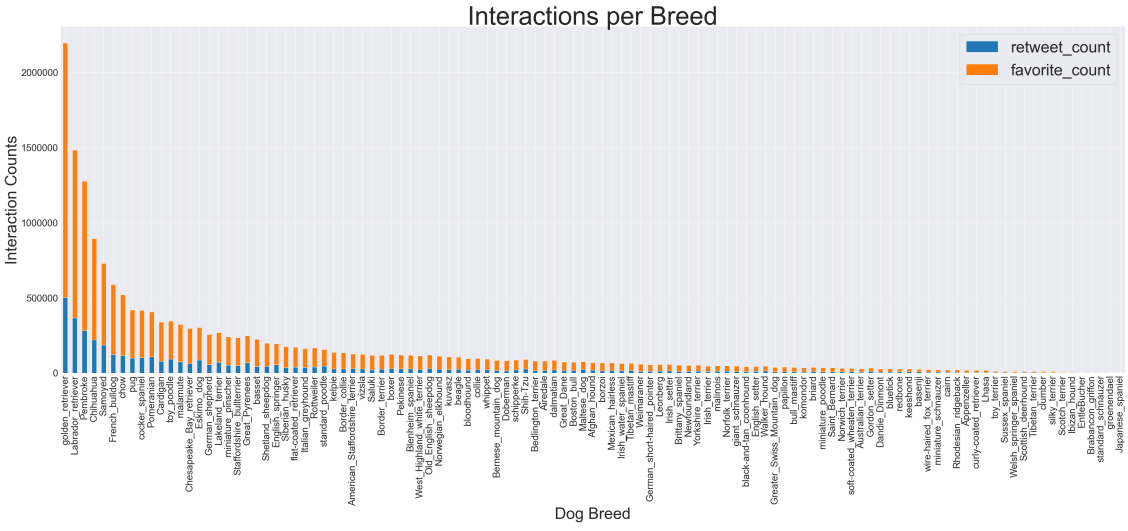
It looks like the users are much more likely to interact with 'doggo' or 'pupper' dogs than the others.

Studying dog breeds and how it affects the user interactions.

[46]:
```
ax = df['p1'].value_counts().plot(kind='bar', figsize=(30,10), fontsize=18)
ax.set_xlabel('Dog Breed', fontsize=30)
ax.set_ylabel('Counts', fontsize=30)
ax.set_title('Count of Breeds', fontsize=40);
```



[47]:
```
temp = df.groupby('p1')[['retweet_count', 'favorite_count']].sum().
↪sort_values('favorite_count', ascending = False)
ax = temp.plot(kind='bar', stacked=True, fontsize=30,  figsize=(60,20))
ax.legend(fontsize=50);
ax.ticklabel_format(axis='y', style='plain')
ax.set_xlabel('Dog Breed', fontsize=50)
ax.set_ylabel('Interaction Counts', fontsize=50)
ax.set_title('Interactions per Breed', fontsize=80);
```



0.3 Results

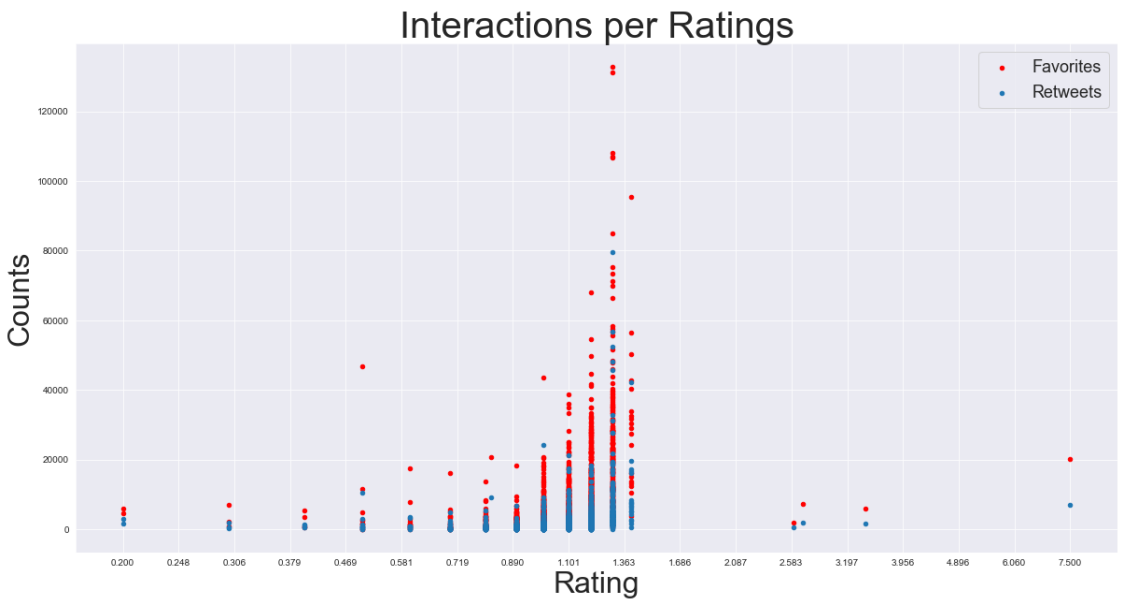
Clearly the 'golden retriever' and the 'labrador retriever' are leading in both, count and user interactions.

Studying user interactions based on rating of the dogs.

[48]:
```
ax = df.plot(kind='scatter', x='rating', y='favorite_count', figsize=(20, 10),␣
↪color='r')
df.plot(x='rating', y='retweet_count', kind='scatter', ax=ax);
ax.set_xlabel('Rating', fontsize=32)
ax.set_ylabel('Counts', fontsize=32)
ax.set_title('Interactions per Ratings', fontsize=40)
ax.set_xscale('log')
ax.set_xticks(np.geomspace(df['rating'].min(), df['rating'].max(), num=␣
↪df['rating'].nunique()))

ax.xaxis.set_major_formatter(matplotlib.ticker.ScalarFormatter())

ax.legend(['Favorites', 'Retweets'], fontsize=18);
```



0.4 Results

The twitter account is known for giving ratings more than 1, normally around 12/10, which is what the users most interact with, but going over the limit with overly high values does NOT bring more interactions with it.

Checking whether the used platform affects user interactions.

[49]:
```
src = df.groupby('source')[['retweet_count', 'favorite_count']].agg(['count',␣
↪'mean', 'median'])
src
```

[49]:

	retweet_count			favorite_count	\
	count	mean	median	count	
source					
TweetDeck	7	1965.000000	919	7	
Twitter Web Client	19	1717.736842	242	19	
Twitter for iPhone	1437	2821.402923	1473	1437	

	mean	median
source		
TweetDeck	5402.142857	3444
Twitter Web Client	4399.157895	559
Twitter for iPhone	9416.806541	4582

Most of the posts are from 'Twitter for iPhone', which surprisingly has higher average interaction than normal.

[ ]: